DOCUMENT RESUME

ED 420 957 EC 306 529

AUTHOR Gersten, Russell; Lloyd, John Wills; Baker, Scott
TITLE Designing High Quality Research in Special Education: Group Experimental Designs.
INSTITUTION Council for Exceptional Children, Reston, VA.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE 1998-00-00
NOTE 48p.; Written "with Lynn Fuchs, Joanna Williams, Sharon Vaughn, Lee Swanson, Susan Osborne, Martha Turlow, Deborah Simmons, and Doug Carrine."
PUB TYPE Opinion Papers (120) -- Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Action Research; Classroom Research; *Disabilities; Elementary Secondary Education; Instructional Effectiveness; *Intervention; *Research Design; *Research Methodology; *Special Education; Statistical Analysis

ABSTRACT

This paper, a result of a series of meetings of researchers, discusses critical issues related to the conduct of high-quality intervention research in special education using experimental and quasi-experimental designs that compare outcomes for different groups of students. It stresses the need to balance design components that satisfy laboratory standards and those that reflect the complexities of real-life classroom teaching. Several controversial issues in group design are addressed, including the importance of clearly defining the nature of the independent variable, the value of conducting replication studies, measuring implementation, the use of quasi-experimental designs, and selecting measures to evaluate intervention efforts. The first section focuses on defining and making the instructional approach operational. The second section considers designs for in-depth understanding of teaching and learning, thereby probing the nature of the independent variable. The third section is on the selection of dependent measures. Also stressed are the importance of researchers reporting effect sizes as well as probability values from statistical tests and the need for researchers to use a set of common measures when researching a given topic. (Contains 79 references.) (DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 957

Designing High Quality Research in Special Education: Group Experimental Designs

Russell Gersten
University of Oregon/ Eugene Research institute
John Wills Lloyd
University of Virginia
Scott Baker
University of Oregon / Eugene Research institute

With
Lynn Fuchs, Joanna Williams, Sharon Vaughn, Lee Swanson
Susan Osborne, Martha Thurlow, Deborah Simmons and Doug Carnine
Edited By Tom Keating, and Sylvia Smith

Preparation of this document was sponsored by a contract from the U.S. Department of Education, Office of Special Education Programs to the Council for Exceptional Children. However, the views expressed here are not necessarily endorsed by either sponsoring agency.

EC 306529

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Setting the Stage

It is indeed ironic that in an era when the public cries out for more information about research-based practices (Gersten & McInerney, 1997; Kornblet, 1997), the number of empirical studies investigating the effectiveness of special education instructional approaches is at one of the lowest levels in 30 years. Scruggs and Mastropieri (1994) provide insights into why the shortage of group research in special education exists:

> When considering intervention research with students with learning disabilities, one is initially struck by the paucity of such research relative to other types of research in learning disabilities. *Why does this relative scarcity of intervention research exist?* Forness and Kavale (1985) argue that early special education research was conducted by psychologists who were more comfortable evaluating psychological characteristics of exceptional populations than evaluating the effectiveness of classroom interventions. As a result, "special education interventions did not evolve as completely as they should" (Forness & Kavale, 1985, p. 7). It is also possible that fewer intervention research studies are conducted because such studies are simply more difficult to design and execute and more costly in terms of necessary resources (pp. 130- 131).

In other words, much of the research on special populations, which has focused heavily on describing their psychological attributes and levels of educational achievement, has tended to avoid research on the effects of interventions because of the difficulties of designing school-based intervention studies.

Yet group experimental designs remain the primary means for assessing whether educational interventions have beneficial effects for students with disabilities. Although other methods of studying change (e.g., single-subject research or qualitative research) can provide valuable insights into the process of change and enhance understanding of facets of teaching and learning, group designs remain the most powerful method for assessing effectiveness.

In contrast, the major goals of qualitative research methods are rich, in-depth descriptions of learning processes and the development of valid interpretations about subtle aspects of teaching and learning – interpretations that then need to be tested using group research designs. Single-subject designs include rigorous experimental methods that can demonstrate causal relationships between variables and discount alternative explanations, but only after extensive replication is it possible to generalize results to students not in the actual studies.

As concern about the effectiveness of special education increases, so too does the field's need for valid reliable evidence about its interventions. Since 1995, the *Council for Exceptional Children*, in conjunction with its Division for Research, and the *Office of Special Education Programs* have convened three meetings of special education researchers to discuss issues related to increasing the quantity and quality of applied experimental group research in classrooms and community learning environments. OSEP and CEC assembled a group of more than 20 researchers, including editors of many of the most prominent special education journals. The precise purpose of these group meetings evolved over time with the course changing somewhat at each of the three meetings.

These discussions produced stimulating dialogue, occasional heated debate, and successively clearer conceptualizations of major issues and roadblocks to the conduct of quality field research for students with disabilities. Some participants argued that the movement over the past 20 years has been to increase the external validity of research findings by conducting more of it in real classrooms or community settings. Unfortunately, this trend has served to compromise the technical standards of group research, at a time when producing credible, valid evidence is particularly important in research that examines the impact of educational interventions.

What can be done to generate stronger and more compelling evidence about effective educational practices for students with disabilities? We concur with OSEP that

maintaining a focus on the intervention research conducted in real school settings is critical, and believe one of the best ways to support this objective is developing consensus on how to design more rigorous, higher quality applied research studies.

The initial research work was to develop standards for improving the quality of group intervention research. One result would have been a set of standards for journal and grant proposal reviewers to use in evaluating the quality of group research. For several reasons, the group decided that developing a more conceptual treatment of critical research design issues, rather than promulgating a list of research standards, would be a more productive tactic.

The most serious reason for not developing a list of standards is that a good design always incorporates balances and compromises. Thus, no single list would be acceptable in all situations. Furthermore, a stringent list might well discourage rather than encourage a resurgence of intervention research. Editors of journals and authors of contemporary essays about intervention research (e.g., Carnine, 1995; Graham & Harris, 1994; Gersten et al., 1998) reinforced our concern about the dwindling number of intervention studies and the potentially discouraging effect of overly stringent standards.

We also were concerned that absolute research standards would not only limit the number of studies conducted, but that the kinds of studies conducted would be constrained, for example, to tightly controlled instructional approaches of short duration. Many factors influence the effectiveness of interventions and many progressive ways are available to examine these factors. Standards may provide guidelines for conducting some types of intervention studies, but would not facilitate the execution of studies that provide rich descriptions of the dynamic process involved in implementing classroom interventions.

One reality of conducting experimental special education research is that, in general, the shorter the length of the study, the more precision there is in attributing

changes in student performance (i.e., the dependent variable) to the instructional intervention (i.e., the independent variable). However, a critical question always must be to ask whether the effects will persist over time – i.e., do the effects last beyond the duration of the study?

As we learn more about instructional approaches that work in real classrooms, our studies have naturally become more complex, and the lengths of interventions have been extended, sometimes to a full school year. Consequently, it is more difficult to conduct tightly controlled experimental research. Further, more research is being conducted by teacher-researchers in an effort to increase the ecological validity of findings. Sometimes too, participating teachers are given discretion in how to implement the principles of teaching and learning that underlie the study. This type of collaborative research challenges traditional concepts of replication (as well as the conventions for reporting procedures that evolved from experimental psychology).

We want to encourage applied research in schools and community learning contexts – research designs that address the realities of school practices. Some traditional group research standards promulgated in textbooks must be adjusted if meaningful research is to be conducted in these settings. Judging how and when to modify research standards developed from laboratory psychology studies without compromising the integrity of designs is critical. We thought that making such issues explicit, and illuminating principles that help create more valid research in school and community settings, would be more valuable than prescribing hard-and-fast rules.

In more formal (if somewhat antiquated) terms, the unavoidable tradeoff between internal and external (ecological) validity must be actively addressed, and we felt a discussion of these issues would be more profitable than identifying standards that, if applied as a whole, could be met only in experimental laboratory research. However, we also thought it was equally important to alert readers to common, predictable pitfalls in design and execution, which can be so severe that the assertions

made by the researchers are not credible (Cooper & Hedges, 1994; Gersten, Baker, Unok Marks, & Smith, 1997; National Center to Improve the Tools of Educators, 1998; Swanson, in press).

For example, an increasing number of quasi-experimental designs are being conducted that use intact classrooms - or even intact schools (Hunt & Goetz, 1997; Slavin & Maddin, 1995, April) - as the means of assigning subjects to different intervention conditions. These studies are developed in response to real-world concerns about the education of students with disabilities. Our goal is to illustrate ways in which these studies can provide credible evidence to other professionals concerning the effectiveness of different instructional conditions.

Essentially, then, the purpose of this paper is to discuss critical issues related to the conduct of high-quality intervention research using experimental and quasi-experimental designs that compare outcomes for different groups of students. We hope to inform new researchers and share the craft knowledge of these issues supplied by the participants in the OSEP and CEC groups. We will articulate how we, as a research community, sensibly negotiate a balance between design components that satisfy laboratory standards and design components that reflect the complexities of real-life classroom teaching - while providing reliable and clear answers to meaningful questions. In doing so, we wish to acknowledge the contributions of participants of our sessions at the OSEP Project Directors' Meetings in 1995 and 1996.

We hope this document will prove useful to people who seek to produce high-quality research in special education.

Overview of Sections

We want to encourage more research on the effects of interventions for groups of students using well designed methods. For decades, group designs have been attacked both by advocates of single-subject research (in the 1970s and early 1980s) and by advocates of qualitative research (in the past decade). The wave of attacks on group experimental designs for studying research on teaching, no matter how well-intentioned and thoughtful (e.g., Kennedy, 1997; Ball, 1995) has taken a toll on what remains our most powerful tool for understanding the effectiveness and impact of instructional interventions and influencing policy.

Designing high-quality experimental comparisons requires sophisticated consideration of research methods. Rather than presenting an exhaustive treatise on every aspect of research design, we offer recommendations about several key aspects, discussing common and thorny issues that persistently recur. Our goal is not to reiterate the principles of experimental research design as found in standard texts. Instead, we hope to "ground" key principles in the realities of current classroom practice.

Although our experience suggests the virtual impossibility of meeting all the criteria for group research discussed in texts, we believe it is possible to design studies in applied settings that are flexible yet rigorous enough to provide valid information on the extent to which theories about teaching, learning, social or cognitive growth lead to instruction that enhances student outcomes.

In this report, we address several controversial areas in group design, including topics on which group members provided interesting insights and perspectives, but failed to reach consensus. It seemed important to present the diverse views on these dilemmas, provide rationales, and explain some of the alternative solutions that the group posed. Among the controversial issues to be discussed are: the importance of clearly defining the nature of the independent variable, the value of conducting replication studies, measuring implementation, the use of quasi-experimental designs,

and selecting measures to evaluate intervention effects.

Much of the skill in conducting high quality research studies is dictated by how well independent variables are conceptualized and operationalized, the sample of students utilized, and the dependent variables selected to assess the impact of the approach. These, thus, become our major organizers for this paper.

Defining and Operationalizing the Instructional Approach

Operationalizing the Independent Variable in the Real World of Classrooms and Community Learning Environments

Precise descriptions of independent variables are crucial to furthering our knowledge base. Most texts treat the operationalization of instructional method as a fairly routine activity, merely checking for "fidelity" (i.e., checking whether teachers are implementing the approach in the fashion the researchers specify). However, as will be seen, the task is far more intricate than one might think (Kennedy, 1997; Kline, Deshler, & Shumaker, 1992). Researchers often have a very good conceptual sense of what they would like to see during instruction, but only "half-formed images" (Kennedy, 1991) of the types of specific actions and behaviors that constitute the implementations, on a day-by-day, minute-to-minute basis.

The difficulty of operationalizing an instructional approach is exacerbated somewhat by efforts to design interventions that work in the real world. For example, teachers are increasingly included in creating and operationalizing the instructional approaches being investigated in studies. This practice tends to result in more flexible specifications of instructional components than might otherwise occur if university researchers were to develop the details of instruction on their own. Also, to increase the prospects that instructional interventions will be applicable to the realities of classroom practice, high degrees of teacher autonomy in making decisions about how to deliver

the intervention are becoming increasingly common.

Nominal Fallacy

It is important to recognize that instructional labels can vary significantly from study to study, and in some cases can be quite misleading . For example, some studies of Classwide Peer Tutoring involve only aspects of decoding and reading fluency. Others Classwide Peer Tutoring studies are essentially reading comprehension studies. Only by serious analysis of what transpires during the lesson is it possible to understand what leads to the specific outcome. And in reality, each study can describe only imperfectly, or allude to, the myriad details that constitute the precise nature of the independent variable. Yet, through techniques such as meta-analysis and other research synthesis methods, we can begin to discern patterns of practices that lead to improved outcomes for students.

In fact, precision in operationalizing the independent variable becomes increasingly important as research syntheses become more common. The purpose of a research synthesis is to "discover the consistencies and account for the variability in *similar-appearing studies*" (Cooper & Hedges, 1994, p. 4, italics added). When the focus in analyzing multiple studies is on similarities and differences in the independent variable (i.e., the instructional intervention), precise descriptions of instruction are critical.

Swanson's (in press) meta-analysis of the impact of various instructional approaches on students with learning disabilities illustrates this well. Swanson concluded that two approaches (direct instruction and strategy instruction approaches) were almost equally effective in enhancing learning. Both approaches had consistent, moderately strong effects, with virtually identical magnitudes (.68 for direct instruction and .72 for strategy instruction). However, Swanson noted great overlap in the way the two constructs were operationalized, and consequently concluded that classifying type of teaching with terms such as *direct instruction* or *strategy instruction* was problematic of a more fine-grained terminology needs to be used.

Regardless of approach, be it direct instruction (Losardo & Bricker, 1994) or strategy instruction **(Kline, F. M., Deshler, D. D., & Schumaker, J. B. (1992).)**, at some point, researchers realize that the exact details of operationalizing the independent variable directly influences which research questions can be answered in a given study. Thus, it is critical to know how the instructional intervention is actually delivered. We will have more to say about implementation issues later.

Several special education researchers, who have conducted research syntheses on an array of topics, have noted these problems with labeling and description of independent variables. For example, in a meta-analysis conducted by Elbaum, Vaughn, Hughes, Moody, and Schumm (1998) many of the interventions were either not described at all, or described so incompletely that it was not possible to describe what occurred during the intervention.

The Gap Between Conceptualization and Execution

Slippage between the conceptualization of a study and its execution is one of the most common problems in applied research, and unfortunately this problem is sometimes so severe that outcomes are uninterpretable. Of course, an important part of any good intervention study is the serious consideration of rival explanations in accounting for student performance. Conducting more applied research studies does not absolve researchers of their responsibility to try to control for confounding variables (e.g., minutes of instruction in treatment and comparison groups or overall teaching effectiveness of teachers assigned to experimental and comparison groups) and to investigate the way variables are operationalized during the study (e.g., degree of support students are provided in their efforts to identify and explicate a theme in a story) (Williams, Brown, Silverstein, & deCani, 1994).

Researchers should provide detailed descriptions of interventions, providing enough information so that they can be replicated. Any in-depth examination of implementation by use of audiotapes, sophisticated observational systems such as Code

for Instructional Structure and Student Academic Response (CISSAR) (Greenwood & Delquadri, 1988) or The Instructional Environment Scale (TIES) (Ysseldyke & Christenson, 1987), will also greatly enhance the quality of the study. Providing actual transcripts of lesson segments and instructional interactions provides rich insights into the true nature of the intervention (e.g., Echevarria, 1995; Fuchs et al., 1997).

A study of beginning reading instruction by Lovett et al. (1994) is a good example of specifying the nature of the independent variable. Although it is a complex study with clear classroom applicability, it furthers our knowledge base on effective reading instruction for students with learning disabilities because theories and their instructional components are operationally defined in a clear fashion. The learning activities resemble the ones reading teachers use, and subtle but important differences in methods are highlighted and systematically linked to outcomes. The goal of such studies is *refinement* of instructional approaches based on empirical investigation. Clearly the components descriptive of instructional interventions are critical to understanding the relationship between intervention and outcome and reporting findings in a useful, replicable fashion.

A Study in Relation to Other Studies: The Importance of Replication

Gage (1997) argues that respect for *the cumulative nature of human knowledge in the field of education* is fundamental. He reflected on those researchers and journalists who in the early 1980s suggested that no useful, generalizable knowledge had come from educational research that could improve teaching or learning. However, by 1997, Gage's (1978) prediction, that data would emerge over time to support empirical studies of teaching and learning using quantitative methods, had "*been resoundingly upheld by the hundreds of meta-analyses that have been reported in the intervening years*" (p. 19). The predictions of cynical journalists and professors, that no replicable patterns would be found, have been proven false.

Gage's (1997) point is that meta-analyses show that

> *many generalizations in education do hold up across many replications with high consistency . . . despite the fact that replications inevitably differ in the persons studied, in the measurement methods used, in the social contexts involved, and in other ways* (emphasis added, p. 19).

In other words, partial replication (as opposed to precise replication) does occur in educational research, and it should be highly valued.

Research in special education has contributed substantially to the knowledge base on effective educational practices. Numerous recent meta-analyses (Scruggs & Mastropieri, 1994; Swanson, in press; Elbaum et al., 1998) have been conducted and have provided confirmations of findings from seminal studies. Replications converge to form a consistent knowledge base that generalizes across student, teacher, and environmental variables (see Forness & Kavale, 1997).

Several aspects of replication studies seem particularly important in special education, and two, in particular, seem worth discussing here.

Independent Replications

First, replication by individuals not invested in developing the independent variable should be a standard practice of special education research. Walberg and Greenberg (1998) point out that a major failure in contemporary education research is the small number of independent replications. A handful of replication studies in special education have been conducted and demonstrate the importance of this underutilized research practice. Hasselbring, Sherwood and Bransford (1986), for example, investigated the effectiveness of videodisc programs they had no role in developing. Similarly, Klingner and Vaughn (1996) have investigated the effects of reciprocal teaching and Simmons, Fuchs, Fuchs, Pate, and Mathes (1994) studied the conditions under which the effects of peer tutoring were enhanced. More recently, Thomas (1998) replicated a vocabulary intervention developed by Beck and her colleagues.

Walberg and Greenberg (1998) discussed the importance of independent replication with extremely popular programs, such as Success for All. Walberg and Greenberg point out that evaluations of Success for All carried out by the developers of the program result in numerical effects on achievement that are among the largest reported in the literature. Independent evaluations, however, result in much more modest effects.

Johnston and Pennypacker (1992) provide a valuable description of the importance of this type of independent, dispassionate replication:

> Regardless of the care with which the original study was done, the credibility of its findings will remain limited until other researchers repeat the experiment and are able to reproduce the original results at least fairly well. Scientists have learned that this is an important step because there are often unidentified factors unique to a particular research program that turn out to be crucial to obtaining the results (p. 245).

A second standard replication practice in special education should be conducting studies to determine which components of a complex instructional approach are critical for achieving an impact on learning or social competence.

In a study using a contrasted groups design, Gersten, Carnine, and Williams (1982) observed instructional interactions of 11 teachers implementing an intensive phonemic approach to teaching beginning reading to at-risk students. They compared observational data gathered on teachers whose students made the most growth in reading with observational data in classrooms where students made minimal growth. The two variables that tended to distinguish most clearly the high growth from the low growth teachers were (a) responding to student errors and problems immediately and (b) maintaining a success rate of at least 85% during the lesson with all students, even those placed in the lowest reading group. These findings were replicated (Gersten, Carnine, Zoref, & Cronin, 1986) with a larger sample of teachers the following year.

Stallings (1980), Leinhardt et al. (1981), and Brophy and Good (1986) have also documented the importance of these two instructional variables in their research- immediate feedback when students make oral reading errors and having students read with high levels of success (especially when they are first learning to read). Unlike replication studies where impartial independent investigation is performed, components analysis warrants study by those with in-depth understanding of the intervention.

Although rewards can be great for conducting studies that are particularly novel, researchers also advance knowledge by conducting studies that complement and extend the current knowledge base. When a research team discusses issues and, ultimately, conceptualizes research questions that emerge inductively from extant research base, it has taken the first, and perhaps most important, step in developing a study. This grounding in literature is important regardless of research paradigm. Researchers who develop research questions may learn that the topic that interests them has already been investigated to some extent. For example, dozens of studies have evaluated whether:

- Delivering skill and content knowledge using principles of direct instruction fosters higher achievement (meta-analysis, White, 1988),
- Teaching students to record whether they are paying attention increases their attention to task (meta-analysis, Lloyd & Landrum, 1990)
- Teaching students comprehension strategies improves their reading achievement (e.g., Wong, Wong, Perry, & Sawatsky, 1986)

An essential part of developing a research question for a replication study is to examine the extant research literature to determine what has been done previously and to identify questions that still need to be answered. Researchers who study related, prior research by conducting a formal literature review in the area find that they can develop a research question of particular benefit even if prior work has been done in the area. They can summarize the extant evidence on their topic, and on the basis of that literature, identify the next few questions that logically should be addressed. It is worth

noting that such a convincing logical sequence of ideas and questions often forms the core of successful proposals for grant funding.

A close relative to developing research questions on the basis of research literature is conceptualizing a series of related studies—what's often called a "replication series"—that examine different aspects of the same higher-order research question. Rather than conducting a set of one-shot studies of a specific intervention ("See, X works!") research teams examine the extent to which "X" works under different circumstances, with different students, teachers, outcome measures, etc.

Assessing The Nature of Instructional Methods and Approaches: Intricacies of Measuring Implementation

Assessing the implementation of educational interventions and approaches has a fascinating history and is one of the most interesting and complex issues in applied educational research. In the 1970s and 1980s, an era of many large-scale evaluations, the importance of assessing the extent to which an educational intervention or approach was actually implemented was stressed in texts and articles. Charters (1974), for example, chastised the field for "evaluation of non-events," i.e., evaluation of programs that for one reason or another were not actually implemented. This led to a rash of studies of how various innovative instructional approaches were actually implemented in classrooms (e.g., Good & Grouws, 1977; Leinhardt, 1977; Stallings, 1975), as well as to the development of sophisticated systems for assessing and understanding implementation, such as the CBAM model developed by Hall and Loucks (1977), which assesses an array of factors related to quality implementation.

To indicate how important implementation is for large-scale intervention research, we refer to a study by Hasselbring et al. (1988) on assessing the impact of a laser disc instructional program in mathematics. These researchers found that, although the program was intended for daily use, many teachers used it only once each week. In other words, these researchers were essentially evaluating a "non-event." In contrast, a

follow-up study by Woodward and Gersten (1992), with careful implementation monitoring, revealed daily use of the program by all of the teachers. Clearly, results from two studies of a similar intervention must be interpreted cautiously in light of marked differences in implementation. In the context of the independent variable, although they may have been nominally similar, they were functionally different.

Many stumbling blocks exist to valid assessment of implementation. Several factors curtailed the major advances in measurement of implementation made in the 1970s and 1980s. The first were drastic budget cuts, which resulted in a reduction of large-scale educational research studies. Implementation research, especially implementation research that requires direct observation of classroom interactions, is expensive. In fact, our experience suggests that assessment of implementation can be as costly as administration of pre- and post-intervention measures. In the next section, we provide a brief overview of the issues and current thinking on the topic of measuring fidelity of implementation and assessing the extent to which the comparison group may also be inadvertently implementing critical components of the intervention.

Measuring fidelity

It is essential that researchers gather and report "core" fidelity of implementation information such as: how much training was provided to the participants, length of lessons, whether critical aspects of teaching were in fact implemented in each classroom, how much time each day was dedicated to the intervention, etc. (See for example, Kelly et al., 1990)

This first method of assessing implementation is *important, but insufficient* to advance the knowledge base in important ways. With this method, a fidelity-of-implementation checklist is developed, and impartial observers rate whether crucial aspects of an intervention are occurring. For example, for class wide peer tutoring, an observer would drop into a classroom and note whether students read in pairs, were awarded points, and used some kind of error correction strategy. For

interventions based on the National Council of Teachers of Mathematics reforms in math, an observer might assess whether students worked in cooperative groups, whether worksheets were used, and whether students were provided with problems that had more than one correct answer.

A major limitation of this type of fidelity of implementation checklist is that the *quality of implementation* is assessed only indirectly. In other words, the more elusive aspects of superior implementation (such as quality of examples used, type of feedback provided) that often are the heart of complex interventions may not be assessed by a checklist.

One reason for the decline in studies of implementation was a convergence of findings (e.g., Cooley & Leinhardt, 1980), that seemed to indicate that generic features of effective instruction transcended any given intervention. Initially, this was an unanticipated finding based on studies that attempted to link degree of implementation with student outcomes. As researchers began to probe in increasing depth the precise features of direct instruction (Gersten et al., 1982), individualized instruction , and other interventions geared to low-achieving students , they found that common features among these models –academic engaged time, continuous monitoring of student progress, quality of feedback provided to students as they encounter difficulties – were as important as the nominal label of the intervention.

However, in the past decade, qualitative studies of implementation, often using discourse analysis, have begun to appear in the literature (e.g., Ball, 1990; Fuchs & Fuchs, 1994; Gersten, 1996; Palincsar & Brown, 1984; S. Williams & Baxter, 1996). These studies use analyses of audiotapes or videotapes to capture the nuances of implementation. This use of selected verbatim transcripts has helped substantially in understanding what really happens when a class uses class wide peer tutoring, content-area sheltered instruction, reciprocal teaching, or NCTM-based math instruction.

Another important point is the time frame for conducting implementation checks. At minimum, implementation checks should occur at the beginning of the study, again a few weeks later to verify corrections, and again midway to late in the study to check for "slippage." When reporting implementation, it is helpful to specify the periods when implementation checks were conducted. Finally, assessing fidelity of implementation allows the research team to check for slippage later in implementation, i.e., unplanned deviations from the intended instructional approach.

Contemporary investigations of implementation have also begun to assess teachers' understandings of the underlying thinking behind the intervention. These investigations have allowed for a deeper level of understanding of how teachers adapt interventions, and the extent to which these adaptations have integrity. In the words of Graham and Harris (1994):

> Intervention integrity expands on the concept of treatment integrity by requiring assessment of the processes of change as related to both intentions and outcomes. . . . Determining the relative contributions of instructional components and the variables responsible for change represents a major challenge. (p. 151)

Interviews with teachers, especially ones that focus on rationales for specific options, will greatly enhance our understanding of the feasibility of the intervention. It will also clarify what we mean by implementation with integrity.

In summary, assessment of implementation is complex, and too often neglected. One major reason is that assessment of implementation is often difficult and expensive to do properly. Also, to date the field has devoted insufficient attention to the topic.

The Nature of the Comparison Group

The following comment about comparison groups illustrates one clear way in which the quality of intervention research can be improved:

> Interventions are best evaluated relative to credible comparison conditions. . . . One way to improve the design of credible alternative conditions is

> communication (and potentially even collaboration) with scientists who are well informed about the alternative interventions. To the extent that intervention researchers perceive studies to be horse races – that are either won or lost relative to other interventions – constructive communication and collaboration with workers representing alternative interventions is unlikely (Pressley & Harris, 1994, p. 197).

One of the least glamorous (and most neglected) aspects of research is describing and assessing the nature of instruction in the comparison group. Yet, to understand what an obtained effect means, one must understand what happened in the comparison classrooms.

This is why implementation of comparison classrooms should also be assessed by members of the research team. At a minimum, researchers should examine comparison classrooms to assess what instructional events are occurring and what professional development and support is provided to teachers. Factors to assess are: access to the curriculum content actually being assessed, time allocated to instruction, and type of grouping used during instruction (Elbaum et al, 1998). Some other questions that should be explored include:

- Did the two conditions differ markedly in the amount of feedback available to the learners?
- How many demonstrations and practice opportunities were provided to each group?
- Were the learners in both groups equally likely to receive personal encouragement for persisting in solving problems?

Perhaps the most serious threats to interpretation are related to what Swanson (in press) labels one of the major problems in special education research: confounding teachers with intervention approach, i.e., classroom by treatment confounds. In particular, we must ensure that the quality of teaching is similar across the two conditions. An ideal means of accomplishing this goal is to "counterbalance" teachers across both treatment and comparison conditions. If counterbalancing teachers and

instructional condition is not possible, one should never rely on a single teacher to deliver an innovative approach, and should develop means to assure readers that teaching quality is basically equivalent across

Defining and Operationalizing the Instructional Approach

Recommendations:

- Avoid the "nominal fallacy" by carefully labeling and describing independent variables.
- Conduct independent replications of intervention research.
- Assessment of implementation needs to be addressed in a serious fashion.
- Carefully document and assess what happens in comparison classrooms or other settings.

the conditions. It is also good to assess "teacher effects" on a post hoc basis, using typical analysis of variance procedures (Dimino, Gersten, Carnine, & Blake, 1990).

Swanson (in press) notes that another serious threat to interpreting treatment effects is that in which intervention studies "stack the treatment condition" with more steps and procedures than the comparison condition.

Designs for Indepth Understanding of Teaching and Learning: Probing the Nature of the Independent Variable

In the past five years, an increasing number of instructional researchers have called for alternatives to traditional experimental or quasi-experimental designs for use in research on teaching and learning. The central problem is that efforts to control, manipulate, and understand a narrow and precisely defined independent variable

rarely result in either a deep understanding of the realities of classroom implementation, nor do they reveal how the independent variable, interacting with other aspects of instruction, contributes to how complex content is learned. One partial solution is to conduct flexible studies that allow for a deeper understanding of what an independent variable might actually look like, given the realities of classrooms, in advance of conducting formal experimental and quasi-experimental studies.

Calls for the increased use of such alternative designs originate from a range of disciplines (e.g., technology, science education, cognitive science, reading comprehension), and the nature of these proposed alternatives is remarkably similar. These studies are increasingly visible in the literature and go by a variety of names: design experiments (Brown, 1992), formative experiments (Newman, 1990; Reinking & Pickle, 1993), or developmental studies (Gersten, Baker & Dimino, 1996; Gersten & Baker, 1997).

We do not envision design experiments as supplanting experimental research. Rather, design experiments may be useful tools for conducting research on newer, less well-defined topics such as technology applications, integration of technology with instruction, and studies of teacher change.

Reinking and Pickle (1993) provide a helpful conceptualization of what a design experiment is and how it unfolds:

> [The] plan for implementation . . . conceived at the beginning is seen as [a] first draft, subject to modification during the experiment. Through systematic investigation, the researcher(s) observe and document factors that inhibit or enhance implementation of the intervention and achievement of the pedagogical goal (p. 264).

In other words, Reinking and Pickle suggest that in design experiments, researchers begin with specifying the desired goal. Then, based on student performance data collected and responses from teachers who are implementing the intervention,

they continually modify the intervention to reach the goal. This is necessary because with less-tested instructional methods, until one is actually immersed in a classroom, it is not possible to select the appropriate length of intervention or the most valid, relevant dependent measures.

It is important to note that advocates of design experiments (Beck et al., 1996; Brown, 1992; Reinking and Pickle, 1993; Richardson & Anders, in press) do not argue for total abandonment of quantitative measures of student learning performance. Rather, they argue for a mix of qualitative and quantitative methods within the context of a design experiment.

Although these designs are still in the early stages of development, we believe they hold promise for those involved in instructional research. The limitations of conventional designs for conducting research on classroom teaching and learning raise significant issues for the entire educational research community.

This point is made in a rather witty fashion by Reinking and Pickle (1993). Their description captures the frustrating experience of many who conduct applied instructional research in classrooms. They begin by describing their proposed study of computer use and literacy and the unanticipated changes and compromises they needed to make for successful implementation:

> As the school year progressed, we found ourselves in a seemingly endless cycle of compromises that threatened the control required in a true experiment . . . Each compromise seemed like a defeat in a war we were quickly losing . . . Our need to maintain control of extraneous variation was a barrier to finding and understanding the most relevant aspects of implementing the intervention and the effects it might have on the educational environment (pp. 266-267).

The problem of maintaining control over extraneous variables has been exacerbated as instructional researchers have extended the length of interventions from days to months and have moved away from easy-to-measure skills such as math

computation and attacked more complex problem-solving skills characteristic of the real world of classrooms.

As researchers have attempted to study changes in teaching, the drawbacks of traditional designs have become apparent. Richardson and Anders (in press) note that studies of implementation "almost invariably take unexpected twists and turns" (p. 25) as researchers examine teachers' adaptations of instructional practices developed by researchers. They note how formal quasi-experimental designs, or studies where the variables are too precisely defined in advance, can inhibit understanding of the process to be investigated. Based on our own experience and reading of research on the change process, we believe this is a valid point.

Richardson and Anders (in press) urge that in studying a complex issue such as teacher change, researchers should concurrently collect both "objective" and more subjective data and be "open to surprises and new understandings in learning about the process and its results" (p. 25). Ideally, research questions, emerging research issues and "results" should be investigated "in a way that allows the reader to continue thinking about the process, data, and consequences" (p. 25).

Brown's (1992) extensive use of instructional interviews to assess students' understandings of scientific content demonstrates the advantages of flexibility in investigating aspects of student learning that occur in new areas of inquiry. She begins by asking a child a series of basic questions dealing with factual or declarative knowledge of key scientific concepts such as the food chain or photosynthesis. If the student is unable to answer, Brown provides examples or prompts. If the student appears to know the concepts, depth of understanding is probed with a series of examples and counterexamples.

Brown (1992) argues that only this in-depth probing through a range of examples has enabled her research team to understand what students really learn about scientific concepts, and which concepts require additional discussion. In Brown's words, these

dynamic assessments "allow us to track not only retention of knowledge, but also how fragile-robust it is and how flexibly it can be applied" (p.159). This information is then used to refine aspects of the independent variable by conducting a formal experiment and to sharpen the sensitivity of dependent variables to assess what is really being learned.

In a similar way, Pressley and El-Dinary (in press) describe how their open-ended design experiments of comprehension strategy instruction permitted them "to construct a far more complete model of comprehension strategies instruction. The insights . . . gained from . . . [the design experiments] permitted the design of a quantitative, comparative study of teacher-implemented comprehension strategies instruction that, we believe, was more realistic than previous studies of the effects of comprehension strategies instruction . . . " (p. 11-12). Interventions based on design experiments tend to be more dynamic and more responsive to the complexities of classroom environments than those developed in isolation at a university or research institute.

We believe that the design experiment can and should be a critical tool in refining innovative instructional practices in real classroom environments and then formally documenting those effects. Design experiments are particularly useful in gaining an in-depth understanding of the relationship between specific aspects of instruction and learning on a variety of performance measures and the nature of effective adaptations for students with disabilities.

Selecting, Describing ,and Assigning Students to Conditions

Identifying and describing the group being studied is central to designing and implementing quality group design. Although research textbooks often describe this process as relatively straight-forward, it is actually quite complex for many reasons that will be explored in this section.

Sample Size

One of the most frequent questions about group-comparison studies is, "How many students do I need in each group?", or some variation such as "Would it be O.K. if I had 12 children in each group?" Researchers conducting studies with special populations are faced with the extraordinary challenge of identifying populations that are sufficiently homogeneous to constitute a group and yet large enough to provide adequate power for group comparisons. Too often, intervention studies with special populations yield nonsignificant results because there are too few participants in each group. As a general rule, more is better. Usually, the old aphorism "20 is plenty" (i.e., 20 students per condition is adequate) still holds. Rarely will samples of 12 or 15 be adequate unless the anticipated effects are strong.

Relative strength of treatments. When a research team is studying an intervention expected to produce substantial effects, the need for large sample sizes decreases. Thus, making a comparison between a powerful and a weak intervention will require fewer students in each condition than when comparing two versions of a powerful treatment with only minor variations between them.

Studies that contrast slight variations of powerful interventions are increasingly being conducted (e.g. Bottge & Hasselbring, 1993; O'Connor & Jenkins, 1995; Graham, MacArthur, & Schwartz, 1995; Fuchs, Fuchs, Kazdan, & Allen, in press). These types of studies require more statistical power than those that ask simpler questions in which experimental treatments are compared to no-treatment controls. Too frequently, researchers forget about this issue when considering an adequate sample size.

An alternative way to increase power is to increase the homogeneity of the groups involved in the study. There are tradeoffs to this technique. With groups of students of quite similar ability (e.g., only students reading between 2.1 and 2.4 on the Woodcock Reading Mastery Test), the variance attributable to extraneous factors will be reduced and thus statistical power increased. However, this technique limits the generalizability of findings. Furthermore, students with similar scores on the Woodcock

Johnson may be quite heterogeneous in terms of other salient variables, such as receptive vocabulary, and task persistence. Thus, rarely is power is increased by restricting the range of the sample.

More thorough sample descriptions. The importance of adequately describing samples increases with the growing emphasis on synthesizing research findings. Swanson (in press) notes, for example, how reading studies involving students with learning disabilities showed larger effects with a sample of students at or below the sixteenth percentile. As previously mentioned, McKinney, Osborne and Schulte (1993) found that attention deficits had a major impact on how well students responded to instructional intervention.

The Research Committee of the Council for Learning Disabilities (Rosenberg et al., 1994) noted that available descriptions of individuals with disabilities in research reports are vague and inconsistent. Inadequate descriptions of participants make it difficult at best, and sometimes impossible, to evaluate research findings or to replicate studies. These problems are particularly acute for studies involving students with learning disabilities because of the complexity and variability of definitions used to determine their eligibility for special education. The committee recommended that the following descriptions be used for describing students with disabilities:

- gender
- age
- race or ethnicity
- level of English language development
- socioeconomic status
- intellectual status of the participants
- achievement levels on standardized tests

The committee recommended that participants be described in narrative or table format. In research with small sample sizes (fewer than 10 participants), a more thorough description is warranted (see Klingner & Vaughn, 1996 as a model).

Correlations between pretest and outcome measures should be routinely

conducted to gain an understanding of factors that help us understand which types of students are most likely to benefit from the intervention within a given sample. For populations as diverse as those typically involved in special education research, these secondary analyses can be extremely important. This is especially true for disability categories such as learning disabilities where comorbidity is common. McKinney et al. (1993) for example found that academic outcomes were quite different for students with learning disabilities depending on the presence or absence of attention problems *even if students began the study with equivalent pretest scores*. Information on subtypes of disabilities and differential impacts of instructional interventions can be extremely valuable to the field. Klingner and Vaughn (1996) conducted secondary analyses to determine which of the English language learners in their LD sample were most likely to benefit from reciprocal teaching.

Random Notes: Assignment Versus Selection, A Major Source of Confusion

Distinguishing between random selection and random assignment is important. The relative importance of the two is sometimes confused, with devastating results . For survey research or demographic studies, random selection is critical. Using random selection, a research team can generalize their results to the population from which they drew their sample, thus addressing a particular study's external validity.

Although *random assignment* of students to treatment and comparison conditions is critical in intervention research, random selection is not. Randomly assigning students to experimental and comparison conditions provides greater certainty that differences between groups on outcome measures are the result of the treatment. This is primarily a matter of internal validity.

One way to achieve comparable groups – and high precision in a study – is to match pairs of students on a variable salient to outcomes in the study (e.g., reading fluency in a reading comprehension study) and then *randomly* assign one member of each pair to each condition (Cook & Campbell, 1979). Note that this procedure is quite

different than finding a "match" for an experimental student after students in the experimental condition have been determined. Matching students on an important variable and then randomly assigning them to treatment and comparison conditions leads to a well-controlled true experiment. Finding a match for a student in the experimental group from a neighboring classroom or school leads to a quasi-experiment, with numerous potential problems outlined in detail by Campbell and Stanley.

Another viable approach is to *stratify* students on a salient variable such as reading or writing ability as measured on standardized or performance measures, and randomly assign within each stratum (i.e., those with low scores, average scores, high scores). Both of these methods (as well as simple random assignment) are legitimate approaches, each with its own strengths and weaknesses.

When random assignment is not possible, quasi-experimental designs may be the only viable alternative. However, validity of inferences drawn from quasi-experiments will always be subject to question (Cook & Campbell, 1979).

Controversies Surrounding Use of Quasi-Experimental Designs in Special Education Research

Since the publication of Campbell and Stanley's (1963) classic monograph, urging those involved in field research to conduct quasi-experiments rather than "true experiments" involving random assignment of students to conditions, researchers have often utilized *quasi-experimental designs* in their intervention research. In true experiments, subjects are randomly assigned to one of the intervention or treatment conditions. In quasi-experiments, researchers often use students from intact classes or schools as the intervention or treatment sample and try to find a relatively comparable group of students from other classes or schools to serve as the comparison sample.

In a quasi-experiment, to explain or account for potential differences between the treatment and comparison groups, researchers typically assess students on a battery of

pretest measures to ensure equivalence. If differences exist, analysis of covariance can be used to adjust statistically for these initial differences. Increasingly, in studies involving three or more data points, sophisticated techniques such as growth curve analysis (Dunst & Trivette, 1994) are used.

However, use of quasi-experimental designs remains controversial, as a recent article by Greeno (1998) notes, and as Campbell himself realized (Campbell and Erlebacher, 1970).

Problems with quasi-experimental designs are particularly severe when pretest differences between treatment and comparison groups exceed one half standard deviation unit on relevant criterion measures. In these cases, it is likely that students in the two groups come from different populations. Of course one can never match or use covariance on every possible variable on which the groups may differ, and it is always possible that an unknown variable differentiating the groups is actually responsible for the posttest results – not the intervention. For this reason quasi-experiments can never really supplant true experiments.

Our concern is that the increased use of quasi-experiments without adequate attention to important features of research design has resulted in many studies that are so weak or compromised that it is unclear that the data support the assertions made by the researchers (e.g., because of clear bias in selecting students for the intervention group or possible large initial differences at pretest.)

Researchers who conduct meta-analyses frequently exclude quasi-experiments unless the researcher demonstrates that no more than one-fourth of a standard deviation separates experimental and comparison groups on salient pretest variables (National Center to Improve the Tools of Educators, 1998) and provides evidence that quality of teaching is not a confound. But of course a quasi-experiment is never going to be an ideal substitute for true experiment, no matter how well designed and conducted, and certainly no matter what the results.

Scruggs and Mastropieri (1994) noted, "Assessment of relative treatment efficacy is extremely difficult from a design that essentially confounds treatments with classrooms" (p. 133). Swanson (in press) noted that confounding treatment with classroom (and teacher) was the major flaw in special education research.

Increasing the Quality of Quasi-Experiments

Because of the limitations of quasi-experiments, some researchers feel we need to encourage increased use of true experiments – where subjects are randomly assigned to conditions – and rely much less on quasi-experiments. Cases of intervention studies with random assignment of students to treatment and high ecological validity are present in the special education literature.

Some members of the OSEP Work Group felt, however, that given the constraints of working in schools and clinics, true experiments were impossible to conduct on a large scale. Others argued that researchers often fail to put sufficient energy into negotiating for random assignment. The Work Group concluded that since quasi-experiments are a way of life for many researchers, standards for conducting quasi-experiments should be more seriously maintained, and the results of well-conducted quasi-experiments considered as serious research.

The group concurred that the first essential standard for quasi-experiments should be adequate pretesting of subjects. Invariably several pretest measures are required to demonstrate comparability between groups. Moreover, these measures must have documented reliability and validity.

Quasi-experiments with no pretest data should not be considered acceptable for publication in journals or for widespread dissemination, unless strong disclaimers are provided. Also, studies in which there is more than a .5 standard deviation unit difference in pretest scores on important variables should be seriously scrutinized before publication in journals, or as evidence supporting the effectiveness of a particular intervention or instructional approach.

Another issue that emerged in the discussions was the fragility of analysis of covariance as a means of "correcting" for initial pretest differences between experimental and comparison samples. Use of analysis of covariance should be limited to cases where initial differences are quite small, i.e., no more than one-half SD unit. In addition, the standard assumptions mentioned in every textbook must be met.

Growth curve analysis can be used in quasi-experiments if more than two testing occasions are included in the design. With sophisticated data analysis procedures such as growth curve analysis, the likelihood of erroneous inferences is somewhat reduced. The reason for this is that with growth curve analysis, it becomes clear when students are from different populations (or from populations that react differentially to the intervention). With analysis of covariance, the assumption is made that all students emanate from the same population. Unless this assumption is true, attempts at statistical control for initial differences are ineffective or invalid (Campbell and Erlebacher, 1970). (Note that growth curve analysis is also an excellent technique for use in true experiments that involve more than two data points.)

For all quasi-experiments, the author should explain how she or he attempted to control for extraneous variables (such as confounding the intervention with teacher effectiveness), include a rationale for why this type of quasi-experimental design was selected, and provide some type of disclaimer. If the author is candid about the design limitations, indicates that the results are largely exploratory, and links the findings to prior research, quasi-experiments can contribute to the accumulated knowledge on a topic. Yet they will never substitute for true experiments.

For quasi-experiments, the responsibility is on the researchers to demonstrate that the effects are due to the intervention, and not to other factors. This burden of proof is much higher than in studies with random assignment.

Designs for Indepth Understanding of Teaching and

Learning: Probing the Nature of the Independent Variable

Recommendations:

- Provide a thorough description of samples.
- Strive for random assignment where possible.
- Provide a rationale for using quasi-experimental designs, including controls for extraneous variables and design limitations.
- Explore the use of alternative designs, such as formative or design experiments.

The Selection of Dependent Measures

Far too often, the weakest part of an intervention study is the quality of the measures used to evaluate the impact of the intervention. Thus, a good deal of the researcher's effort should be devoted to selection and development of dependent measures. In essence, the conclusion of a study depends not only on the quality of the intervention and the nature of the comparison groups, but also on the quality of the measures selected or developed to evaluate intervention effects.

Often intervention researchers spend more time on aspects of the intervention related to instructional procedures, rather than measurement of effects. Although it is understandable that many educators would rather teach than test, and few students enjoy being tested, creating tests of unknown validity invariably weakens the power of a study (i.e., reduces the chances of documenting the intervention's effectiveness), and limits the potential for syntheses because of a lack of common measures.

The importance of selecting, developing, and refining dependent measures to address the full array of research questions in a study is critical in high-quality research. It is usually valuable to use multiple measures in a study, given the fact that any measure is necessarily incomplete and imperfect, and no one measure can represent all

or even most of the important phenomena that an intervention might affect. Any one measure can assess only a facet of the construct of interest, and therefore that measure is necessarily narrow or restricted. That is why research teams should choose several measures to tap various facets of a single construct. However, it is also important when using multiple dependent measures to make sure that appropriate statistical analyses are used to avoid inflating the possibility of finding significant effects.

Note, too, that in many intervention studies researchers may assess more than one construct. For example, in a study examining the effects of socially-mediated instruction on reading outcomes, a team may want to assess differences in terms of: reading fluency, reading comprehension, and social relations.

For each of these constructs, it likely will require more than one measure to validly assess the construct. With reading comprehension, for example, it may be important to ascertain whether the experimental condition affects both literal and inferential comprehension. For social relations, it may be important to ascertain whether the effects are specific to general social standing, as well as close personal friendships.

A test of reading comprehension may require students to (a) read a sentence, paragraph, or story silently or orally and write or say answers to multiple-choice, short answer, or essay questions; (b) orally or silently read sentences or passages that contain blanks and restore those blanks with semantically correct words either orally or in writing; or (c) read paragraphs or stories and write or tell summaries of the content read. Although each type of measure taps some dimension of reading comprehension, it should be clear that the testing requirements differ dramatically, and student performance—and assessed treatment effects—may vary accordingly.

Guidelines for Selecting Measures

Most of the constructs we use in educational research—mathematical problem solving, expressive writing ability, phonemic awareness, self-esteem—are at best,

loosely defined. For these reasons, multiple measures are always a necessity as is a clear rationale of how a particular construct is being used in the study. It is not surprising that there is an art to selecting and developing measures.

Advantages of Using Multiple Measures

One of the foremost challenges to a research team is to select, from available achievement tests, those measures that are well-aligned with the substance of the intervention and that will be sensitive to, or register, treatment effects if they occur. The second—and often competing—challenge, however, is for the researcher to select or develop achievement measures that are sufficiently broad and robust to (a) avoid criticism for teaching to the test through the specific intervention and (b) demonstrate that generalizable skills have been successfully taught through the treatment.

An intervention may have adverse effects or additional benefits that a researcher should attempt to identify by employing multiple dependent measures. For example, exposure to well-designed instruction in solving problems involving fractions may lead to enhanced conceptual understanding of fractions. Similarly, a well implemented conceptual approach to teaching students with disabilities about fractions may, as a side effect, lead to enhanced ability in computation and word problem solving. Only by using a broad array of measures can we begin to address these issues empirically. Finally, the benefits of an intervention are clearer if its effects converge across multiple sources.

Over-Reliance On Closely Aligned Dependent Measures

By and large, intervention effects are stronger to the extent that measures are aligned to specific objectives of the intervention. There are several reasons for including some closely aligned measures in a study. The first is that it makes sense to measure the extent to which students learn exactly what they are taught. This can be important formative information for researchers. For example, can students who have been taught to use story grammar questions, actually answer story grammar questions? Can

students who recently completed a unit on addition and subtraction of fractions, add and subtract fractions?

Although one component of a measurement battery may be biased in favor of the experimental intervention, the data on that measure still provide critical information. If an important variable aligns more closely with one intervention, being clear about this bias is important. Rather than omitting the evidence, we think it is better to provide the data and include appropriate caveats.

For example, Fuchs and Fuchs (1994) describe data from one of their studies involving teacher use of story grammar questions. They note that use of story grammar questions did not have an impact on student's oral reading fluency. But for measures that were more closely related to the intervention, like the Stanford Achievement Test, which requires students to read passages and respond to multiple-choice questions, the effect size was a moderate .37, favoring those students whose teachers incorporated more story grammar comprehension questions into instruction. For the retell task, the measure most closely aligned to the intervention, the effect size was strongest, .67.

Other researchers stress the importance of selecting an array of measures that are not heavily biased toward the intervention that they themselves have developed. Swanson (in press) noted that treatment effects were stronger on experimenter-developed measures than on standardized measures of the same construct. For example, if researchers are investigating story grammar, their comprehension assessment is quite likely to have several measures of story grammar knowledge and use whereas other comprehension measures—which include a broader array of comprehension questions—almost invariably show effects less favorable to the experimental intervention. In the view of many researchers, the ultimate goal is to build broad competencies in students, and too often, we tap very small aspects of these abilities with limited experimenter-developed measures. Conversely, when studies show clear effects on broad-spectrum tests (i.e., a well-standardized achievement

measure), without having explicitly taught the content of the test, we should place substantial weight on the importance of the findings.

It is particularly important that researchers gear the measures they select toward the objectives of the study—which may be to increase reading comprehension, i.e. increasing students' ability to read with understanding—rather than narrow aspects of the intervention itself. Such an approach promotes greater understanding of the phenomena at hand and permits other researchers to ask more intelligent and helpful questions in subsequent research.

On a more mundane, though more practical, note, researchers often err in two different ways in the selection of measures. Some researchers exclusively use "off the shelf," commonly available measures such as standardized achievement tests or well-known self-concept measures such as the Harter scales. Exclusive reliance on these published standardized measures may yield data that are insensitive to the effects of the intervention. Other researchers err in the other direction, relying only on researcher-developed measures, often of unknown validity.

It would benefit the field to make clear the relationships between experimenter-developed measures and those measures that are familiar to readers (such as standardized achievement tests or commonly used measures of social behavior). This should be done both empirically and conceptually. By including well known assessments to supplement those developed by the researcher, the potential for bias is reduced. Presenting intercorrelations between measures gives readers a sense of the extent of overlap, and concurrent validity of the measures. Researchers should include discussions of construct validity, using their own data and data from other relevant research.

The fate of any study is in large part due to the quality of the measures developed and selected. We strongly recommend that pilot research be used to refine measures and that piloting procedures and psychometric characteristics be reported for

all measures used. There is a tendency to use our best passages (in reading research) and our best problems (in math or science research) for the teaching and save the weaker ones for testing. After all, we want the intervention to be the best possible, and to use the best materials possible to obtain the desired effects. Yet this pattern is unwise in group contrast research. In essence, the findings depend on the quality of the individual items selected for assessment. Consequently, using weaker items (passages or problems) or items of unknown quality on assessments, while saving the better ones for instruction, is ill-advised.

Selection of Measures

Recommendations

- Seek a balance between global competence and specific skill measurements.
- Select some measures that are non-biased toward intervention.
- Ensure that not all measures are experimenter-developed; that some have been validated in prior research.
- Select measures that are technically adequate (include reliability and validity).

Although conventional psychometric indices have their limitations, we believe they are often under-utilized in special education research. In particular, estimation of *internal consistency reliability* (often referred to by technical names such as coefficient alpha or Cronbach's alpha) is a critical aspect of a research study. It is, however, one that is often neglected, even in published research articles, even though coefficient alpha reliability is easy to compute with common statistical packages. Internal consistency reliability programs help us understand how well the cluster of items on a test "fit" together, i.e., how well performance on one item predicts performance on

another. Analysis help the researcher locate items that don't fit well, i.e., items with a weak item-to-total correlation. Revisiting these items and either revising them or dropping them can improve the statistical power, and ultimately the quality of the study, if coefficient alpha is conducted on a pilot sample prior to conducting the study.

Research teams may also want to test new or labor-intensive measures such as think-alouds, and retellings, on a small random subsample of students engaged in the study. This is preferable to omitting these measures altogether, and the findings, although exploratory in nature, can help inform future research.

Next Steps

There are two areas on which there was broad consensus for improving the quality of research and our ability to synthesize research. The first was that researchers should routinely report *effect sizes* as well as probability values from statistical tests. An excellent resource for a conceptual treatment of quantitative research syntheses, and for procedures for the calculation of effect sizes, is Cooper and Hedges (1994).

Standard reporting formats should allow researchers to read a given study and extract the necessary information for classifying critical variables regarding the nature of the study and the participants. The study should also provide the necessary information to calculate effect sizes for each relevant dependent measure. Conveying information for the calculation of effect sizes is usually done best in tables displaying descriptive data, which normally include all relevant pretest and posttest mean scores and their standard deviations. Although the most frequent meta-analytic syntheses use effect size calculations on posttest scores, it is also acceptable to calculate posttest effect sizes after adjusting for pretest differences, especially when quasi-experimental designs are used. For this reason, it is important to also report pretest measures and standard deviations for experimental and comparison groups.

A second area of agreement was on the need for researchers to use a set of common measures when researching a given topic. These measures would be selected

by a group of researchers with expertise in that field. They would typically include standardized measures, but might also include innovative measures that have been successfully used in research. They need not be limited to paper and pencil tests, and could include for example, direct observation techniques, measures of oral reading following standardized procedures, theoretically compelling measures such as rapid automatic naming (Denckla & Rudel, 1976), or performance measures using state-of-the art scoring methodology.

This would not preclude the use of additional experimenter-developed measures to study specific aspects of the intervention. But routine use of the small set of core measures in areas such as oral reading, social competence, reading comprehension, task persistence would enhance the process of research synthesis and integration of findings. We strongly urge agencies such as OSEP to specifically request (and within the course of two years) mandate such a process.

Afterword

Our hope is to provide guidance for special education researchers about how to conduct high quality experimental studies using group-contrast methods. To that end, we consulted with literally dozens of colleagues and identified several critical ways in which research teams can conduct high quality studies. We do not offer these recommendations as an exhaustive set of guidelines but rather as suggestions for those who wish to use these sorts of designs and are predisposed to pursuing methods that enhance the quality of findings.

Our recommendations center on familiar issues—selecting research questions and conceptualizing their relation with other studies, assigning participants to groups, choosing and describing measures, describing participants and conditions, and so forth. Obviously, those researchers who are interested can learn more about these issues by consulting standard textbooks on research design. However, we have presented these concepts with particular reference to special education intervention research and how

those issues present special challenges to those of us engaged in studying ways to provide better services to students with disabilities, their teachers, and their parents.

References

Ball, D. L. (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis, 12*(3), 247-259.

Ball, D. L. (1995). Blurring the boundaries of research and practice. *Remedial and Special Education, 16*, 354-364.

Beck, I. L., McKeown, M. G., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text. *The Elementary School Journal, 96*, 385-414.

Bottge, B., & Hasselbring, T. S. (1993). A comparison of two approaches for teaching complex, authentic mathematics problems to adolescents in remedial math classes. *Exceptional Children, 59*, 556-566.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *The third handbook of research on teaching* (pp. 328-375). New York: MacMillan.

Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of Learning Sciences, 2* (2), 141-178.

Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Eds.), *Compensatory education: A national debate*. New York: Brunner/Mazel.

Campbell, D. T., & Stanley, J. C.(1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally.

Carnine, D. (1995). The professional context for collaboration and collaborative research. *Remedial and Special Education, 16* (6), 368-371.

Charters, W. W., Jr., & Jones, J. E. (1974). *On neglect of the independent variable in program evaluation*. Project MITT.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally Publishing.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study.

Education Evaluation and Policy Analysis, 2, 7-25.

Cooper , H. & Hedges, L. (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Denckla, M. B. , & Rudel, R. (1976). Rapid 'automatized' naming (RAN): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14,* 471-479.

Dimino, J., Gersten, R., Carnine, D., & Blake, G. (1990). Story grammar: An approach for promoting at-risk secondary students' comprehension of literature. *Elementary School Journal, 91,* 19-32.

Dunst, C. J., & Trivette, C. M. (1994). Methodological considerations and strategies for studying the long-term effects of early intervention. In S. L. Friedman & H. C. Haywood (Eds.), *Developmental follow-up: Concepts, domains, and methods* . New York: Academic Press.

Echevarria, J. (1995). Interactive reading instruction: A comparison of proximal and distal effects of instructional conversations. *Exceptional Children, 61,* 536-552.

Elbaum, B. E., Vaughn, S., Hughes, M. T., Moody, S., & Schumm, J. S. (1998). *The effect of instructional grouping format on the reading outcomes of students with disabilities: A meta-analytic review.* Miami, FL: University of Miami.

Forness, S., & Kavale, K. A.(1985). De-psychologizing special education. In R. B. Rutherford & C. M. Nelson (Eds.), *Severe behavior disorders of children and youth* (pp. 2-14). Boston: College-Hill Press.

Forness, S. R., & Kavale, K. A. (1997, July/ August). Mega-analysis of meta-analysis. *Teaching Exceptional Children,* p. 4-9.

Fuchs, L. S. & Fuchs, D. (1994). Academic assessment and instrumentation. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 233-245). New York: Springer-Verlag.

Fuchs, D., Fuchs, L. S., Mathes, P. H., & Simmons, D. C. (1997). Peer-assisted strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34,* 174-206.

Fuchs, L. S., Fuchs, D., Kazdan, S. R., & Allen, S. (in press). Effects of peer-assisted learning strategies in reading with and without training in elaborated help giving. *Elementary School Journal.*

Gage, N. L. (1978). *The scientific basis of the art of teaching.* New York: Teachers

College Press.

Gage, N. L. (1997). "The vision thing": Educational research and AERA in the 21st century part I: Competing visions of what educational researchers should do. *Educational Researcher, 26*(4), 18-21.

Garcia, E. (1992). Analysis of literacy enhancement for middle school Hispanic students through curriculum integration. *The Journal of Educational Issues of Language Minority Students, 12*, 131-145.

Gersten, R. (1996). Literacy instruction for language-minority students: The transition years. *Elementary School Journal, 96*, 227-244.

Gersten, R., & Baker, S. (1997). *Design Experiments, Formative Experiments and Developmental Studies: The Changing Face of Instructional Research in Special Education* (Technical Report 98-1). Eugene, OR: Eugene Research Institute.

Gersten, R., Baker, S., & Dimino, J. (1996). *Conceptual Approaches to Teaching History to Students with Learning Disabilities in Integrated Settings.* Grant Proposal Submitted to the US Department of Education Programs. Eugene, OR: Eugene Research Institute.

Gersten, R., Baker, S., Unok Marks, S., & Smith, S. B. (1997). Integrative synthesis of the empirical and professional knowledge bases on effective instruction for English-language learners with learning disabilities and those at-risk for school failure. (U.S. Department of Education Grant - HO23E50013).

Gersten, R., Carnine, D., & Williams, P. (1982). Measuring implementation of a structured educational model in an urban setting: An observational approach. *Educational Evaluation and Policy Analysis, 4*, 67-79.

Gersten, R., Carnine, D., Zoref, L., & Cronin, D. (1986). A multifaceted study of change in seven inner city schools. *Elementary School Journal, 86* (3), 257-276. Focus on behavior analysis in education. Columbus, OH: Charles Merrill.

Gersten, R., & McInerney, M. (1997, April). *What parents, teachers and researchers view as critical issues in special education research.* Paper presented at CEC, Salt Lake City, Utah.

Gersten, R., Williams, J., Fuchs, L., Baker, S., Kopenhaver, D., Spadorcia, S., & Harrison, M. (1998). *Improving Reading Comprehension for Children with Disabilities: A Review of Research. Final Report.* U. S. Department of Education Contract HS 9217001.

Good, T. L., & Grouws, D. A. (1977). Teaching effects: A process product study in

fourth grade mathematics classrooms. *Journal of Teacher Education, 28* (3),49-54.

Graham, S., & Harris, K. (1994). Cognitive strategy instruction: Methodological issues and guidelines in conducting research. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 146-161). New York: Springer-Verlag.

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87*, 230-240.

Greeno, J. G., & Group, M. S. M. T. A. P. (1998). The situativity of knowing, learning, and research. *American psychologist, 53*, 5-26.

Greenwood, C. R., & Delquadri, J. (1988). Code for instructional structure and student academic response (CISSAR). In M. Hersen & A. S. Bellack (Eds.), *Dictionary of behavioral assessment* (pp. 120-122). New York: Pergamon.

Hall, G. E., & Loucks, S. (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal, 14*, 264-276.

Hasselbring, T., Sherwood, R., Bransford, J., Fleenor, K., Griffith, D., & Goin, L. (1988). Evaluation of a level-one instructional videodisc program. *Journal of Educational Technology Systems, 16*, 151-169.

Hasselbring, T., Sherwood, B., Bransford, J. (1986). *An evaluation of the* Mastering Fractions *level-one instructional videodisc program*. Nashville: George Peabody College of Vanderbilt University, The Learning Technology Center.

Hunt, P., & Goetz, L. (1997). Research on inclusive educational programs, practices, and outcomes for students with severe disabilities. *The Journal of Special Education, 31*, 3-29.

Johnston, J. M., & Pennypacker, H. S. (1992). *Strategies and tactics of human behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kelly, B., Gersten, R., & Carnine, D. (1990). Student error patterns as a function of curriculum design. *Journal of Learning Disabilities, 23*, 23-32.

Kennedy, M. M. (1991). Following through on follow through. In E. A. Ramp & C. S. Pederson (Eds.), *Follow through: program and policy issues* . Washington, D.C.: Office of Education Research and Improvement, U.S. Department of Education.

Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher, 26*(7), 4-12.

Kline, F. M., Deshler, D. D., & Schumaker, J. B. (1992). Implementing learning strategy instruction in class settings: A research perspective. In M. Pressley, K. R. Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 361-406). San Diego: Academic Press.

Klingner, J. K., & Vaughn, S. (1996). Reciprocal teaching of reading comprehension strategies for students with learning disabilities who use English as a second language. *Elementary School Journal, 96*, 275-293.

Kornblet, A. (1997). *Which research findings are considered most critical by parents and teachers?* Paper presented at the Council for Exceptional Children Annual Convention.

Leinhardt, G. (1977). Program evaluation: An empirical study of individualized instruction. *American Educational Research Journal, 14*, 277-293.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal, 18*, 343-361.

Lloyd, J. W., & Landrum, T. J. (1990). Self-recording of attending to task: Treatment components and generalization of effects. In T. E. Scruggs & B. Y. L. Wong (Eds.), *Intervention research in learning disabilities* (pp. 235-262). New York: Springer-Verlag.

Losardo, A., & Bricker, D. (1994). A comparison study: Activity-based intervention and direct instruction. *American Journal on Mental Retardation, 98*, 744-765.

Lovett, M. H., Borden, S. H., DeLuca, T., Lacerenza, L., Benson, N. J., & Brackstone, D. (1994). Treating the core deficits of developmental dyslexia: Evidence of transfer of learning after phonologically and strategy based reading training programs. *Developmental Psychology, 30*, 805-822.

McKinney, J. D., Osborne, S. S., & Schulte, A. C. (1993). Academic consequences of learning disability: Longitudinal prediction of outcomes at 11 years of age. *Learning Disabilities Research and Practice, 8*, 19-27.

National Center to Improve the Tools of Educators (1998). *Evaluation of research on educational approaches (EREA).* Manuscript in preparation, University of Oregon.

Newman, D. (1990). Opportunities for research on the organizational impact of school computers. *Educational Researcher, 19*(3), 8-13.

O'Connor, R. E., & Jenkins, J. R. (1995). Improving the generalization of sound/symbol knowledge: Teaching spelling to kindergarten children with disabilities. *Journal of Special Education, 29*, 255-275.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Pressley & El-Dinary (in press). What we know about translating comprehension strategies into practice. *Learning Disability Quarterly.*

Pressley, M., & Harris, K. R. (1994). Increasing the quality of educational intervention research. *Educational Psychology Review, 6*(3), 191-208.

Reinking, D., & Pickle, J. M. (1993). Using a formative experiment to study how computers affect reading and writing in classrooms. In C. Z. Kinzer & D. J. Leu, (Eds.), *Examining central issues in literacy research, theory, and practice* (pp. 263-270). Chicago: National Reading Conference.

Richardson, V. and Anders, P. (in press). A view from across the Grand Canyon. *Learning Disability Quarterly.*

Rosenberg, M. S., Bott, D., Majsterek, D., Chiang, B., Simmons, D., Gartland, D., Wesson, C., Fraham, S., Smith-Myles, B., Miller, M., Swanson, H. L., Bender, W., Rivera, D., & Wilson, R. (1994). Minimum standards for the description of participants in learning disabilities research. *Remedial and Special Education, 15*(1), 56-59.

Scruggs, T. E. & Mastropieri, M. A. (1994). Issues in conducting intervention research: Secondary students. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 130-145). New York: Springer-Verlag.

Simmons, D., Fuchs, D., Fuchs, L. S., Pate, J., & Mathes, P. (1994). Importance of instructional complexity and role reciprocity to classwide peer tutoring. *Learning Disabilities Research and Practice, 9*, 203-212.

Slavin, R., & Madden, N. (1995, April). *Effects of Success for All on the achievement of English language learners.* Paper presented at the annual meeting for the American Educational Research Association, San Francisco.

Stallings, J. (1975). *Follow through program classroom observation evaluation.* Menlo Park, CA: Stanford Research Institute.

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on

task. *Educational Leadership, 9*(11), 11-16.

Swanson, H. L. (in press). What instruction works for students with learning disabilities? Results from a meta-analysis of intervention studies. In R. Gersten, E. Schiller, J. Schumm, & S. Vaughn (Eds.), *Issues and research in special education*. Mahwah, NJ: Erlbaum.

Thomas, C. L. (1998). *The effects of three levels of curricular modifications on the vocabulary knowledge and comprehension of regular education students and students with learning disabilities in content-area classrooms*. Unpublished doctoral dissertation, University of Oregon, Eugene.

Walberg, H. J., & Greenberg, R. C. (1998, April 8). The diogenes factor: Why it's hard to get an unbiased view of programs like 'Success for All'. *Education Weekly*, p. 52.

White, W. A. T. (1988). A meta-analysis of the effects of direct instruction in special education. *Education and Treatment of Children, 11*, 364-374.

Williams, S. R., & Baxter, J. A. (1996). Dilemmas of discourse-oriented teaching in one middle school mathematics classroom. *Elementary School Journal, 97*, 21-38.

Williams, J. P., Brown, L. G., Silverstein, A. K., & deCani, J. S. (1994). An instructional program in comprehension of narrative themes for adolescents with learning disabilities. *Learning Disability Quarterly, 17*, 205-221.

Wong, B. Y. L., Wong, R., Perry, N., & Sawatsky, D. L. (1986). The efficacy of a self-questioning summarization strategy for use by underachievers and learning disabled adolescents in social studies. *Learning Disabilities Focus, 2*(1), 20-35.

Woodward, J., & Gersten, R. (1992). Innovative technology for secondary learning disabled students: A multi-faceted study of implementation. *Exceptional Children, 58*, 407-421.

Ysseldyke, J. E. & Christenson, S. L. (1987). *The instructional environment scale: A comprehensive methodology for assessing an individual students' instruction*. Austin, TX: Pro-Ed.